A quasi-Newton method for total variation regularization of images corrupted by non-Gaussian noise

Rick Chartrand, *Member, IEEE,* and Valentina Staneva
Los Alamos National Laboratory
EDICS: RST-DNOI

***Abstract*— Our aim is to obtain efficient algorithms for image regularization optimized for removing different types of noise. To accomplish this, we combine total variation regularization with a noise-specific way to measure the fidelity between the noisy image and the reconstruction. We find a minimum of the resulting functional with a quasi-Newton method, which converges faster than the common method of gradient descent. As examples we consider Poisson noise and impulse noise. We prove convergence of the algorithm for a large class of fidelity terms.**

I. Introduction and background

Our work has been motivated by the promising results of total variation regularization used for images corrupted by Poisson noise. It is shown in [1] that having a data fidelity term reflecting the noise characteristics of the image provides a better image reconstruction. This leads us to consider total variation algorithms with a general data fidelity term. The method of gradient descent is appropriate in this case, since it allows to minimize easily a large class of functionals. Unfortunately, the convergence rate of this algorithm is very slow. Instead, we generalize the fixed-point method of Vogel and Oman [2], which can also be cast as a quasi-Newton method. In the case of the L^2 data fidelity term, the algorithm has a linear convergence rate [3].

A. Total variation regularization

The problem of denoising an image is an ill-posed inverse problem often solved by means of regularization. We consider the variational approach, where the denoised image is the solution of an optimization problem of the following form:

$$\min_u F(u) = \int R(u) + \lambda \int D(u,d). \tag{1}$$

Here, d is the image to be denoised; $D(u,d)$ is the data fidelity term, which measures the dissimilarity between d and the reconstructed image, u; $R(u)$ is the regularization term in which prior knowledge or assumptions about the solution are enforced; and λ is the regularization parameter that balances the relative effect of the two terms. The regularization term we will use is total variation, $R(u) = |\nabla u|$, for its ability to preserve edges in images. It was first proposed in this context by Rudin, Osher and Fatemi [4], together with a data fidelity term (in effect) of $D(u,d) = |u-d|^2$.

Address: Theoretical Division, T-7, MS B284, Los Alamos National Laboratory, Los Alamos, NM 87544, USA. Phone: +1 +505 667-8093. Fax: +1 +505 665-5757. E-mail: rickc@lanl.gov.

Address: Theoretical Division, T-7, MS B284, Los Alamos National Laboratory, Los Alamos, NM 87544, USA. Phone: +1 +505 606-2159. Fax: +1 +505 665-5757. E-mail: vstaneva@lanl.gov.

II. Noise-based data fidelity terms

A data fidelity term is an appropriate measure of dissimilarity between a noisy image and the reconstruction. Knowledge of the noise can help define this measure. For example, it is shown in [5] using probability arguments that the L^2-norm data fidelity term, $D(u,d) = |u-d|^2$, is most appropriate for removing additive, Gaussian noise. The L^2-norm often works for other types of noise. For Poisson noise, however, the noise is signal dependent; the amount of noise increases with the image intensity. Removing this noise without losing image features requires the amount of regularization to vary spatially. As shown in [1], the ideally suited data fidelity term for this is $D(u,d) = u - d\log u$. We will see in Section V examples of the use of this term, computed using the algorithm of Section III.

The Poisson data fidelity term was used with total variation regularization in the context of positron emission tomography by Jonsson, Huang, and Chan [6].

Another example of non-Gaussian noise is impulse noise, in which a random portion of the pixels are corrupted. A particular case is salt-and-pepper noise, where the corrupted pixels have equal probability of having zero or full intensity. Applying the arguments of [5], [1] yields the following heuristic analysis. We can consider our task to be to find the reconstruction u that maximizes the conditional probability $P(u|d)$. Applying Bayes's Law, we can rewrite this as

$$P(u|d) = \frac{P(u)P(d|u)}{P(d)}. \tag{2}$$

Taking negative logarithms, we see that we wish to find the u that minimizes $-\log P(u) - \log P(d|u)$. Total variation regularization arises from the choice of the prior

$$P(u) = e^{-\beta \int |\nabla u|}, \tag{3}$$

for some constant β; this and other constants will determine the corresponding value of λ. The likelihood $P(d|u)$ follows from the noise model. In the case of salt-and-pepper noise, there is q, $0 \le q < 1$, such that for each pixel i, the probability

is as follows:

$$P(d_i|u_i) = \begin{cases} 1-q, & d_i = u_i \\ q/2, & d_i = 0 \\ q/2, & d_i = 1 \end{cases}, \tag{4}$$

where we have assumed that the intensity values lie in $[0,1]$. Assuming independence, we get roughly:

$$\begin{aligned} P(d|u) &= (q/2)^{|\{i:d_i \neq u_i\}|}(1-q)^{N-|\{i:d_i \neq u_i\}|} \\ &= (q/2)^{\|u-d\|_0}(1-q)^{N-\|u-d\|_0} \end{aligned} \tag{5}$$

(this fails to account for where u is 0 or 1 and a "corrupted" pixel has the same value). Here N is the number of pixels and $\|\cdot\|_0$ the L^0 norm, which is simply the cardinality of the support (and not actually a norm). Taking negative logarithms, we obtain

$$-\log P(d|u) = \|u-d\|_0 \log\big((q/2(1-q)\big) + N\log(1-q). \tag{6}$$

The result is that the reconstruction should be a solution to

$$\min_u F(u) = \int |\nabla u| + \lambda \|u-d\|_0, \tag{7}$$

where λ depends on β and q. In this paper, we only wish to consider convex functionals, so we will approximate the L^0 norm by the L^1 norm, and replace (7) with

$$\min_u F(u) = \int |\nabla u| + \lambda \int |u-d|. \tag{8}$$

The L^1 data fidelity term was introduced by Nikolova [7] for total variation regularization of images with impulse noise; see also [8].

III. A quasi-Newton iteration

Now we turn to our method for computing minimizers of the regularization functionals. We begin with the example of Poisson noise, where we seek to solve the following optimization problem:

$$\min_u F(u) = \int |\nabla u| + \lambda \int \big(u - d\log(u)\big). \tag{9}$$

According to [1], this minimum exists and is unique. We discretize the problem with a uniform, rectangular grid with spacing Δx. We consider u and d to be in vectorized form: if the images are of size $m \times n$, then u and d are vectors of length $N = mn$. Let D_x, D_y be the matrices representing the finite-difference approximations of differentiation with respect to x and y. To be specific, we use forward differencing with Neumann boundary conditions. Thus, $|\nabla u|$ becomes $\sum_{i=1}^N \sqrt{(D_x u)_i^2 + (D_y u)_i^2}$. Since this quantity is not differentiable, we add a small constant ϵ:

$$F(u) = \sum_{i=1}^N \sqrt{(D_x u)_i^2 + (D_y u)_i^2 + \epsilon} + \lambda \sum_{i=1}^N \big(u_i - d_i \log(u_i)\big). \tag{10}$$

First, we consider Newton's method. For that it is required to calculate the gradient and the Hessian of the functional. Let Q_u be the diagonal matrix with entries $\big((D_x u)_i^2 + (D_y u)_i^2 + \epsilon\big)^{-1/2}$, and let $L_u = D_x^T Q_u D_x + D_y^T Q_u D_y$. Then the first two derivatives of (10) can be represented in the following way:

$$\nabla F(u) = L_u u + \lambda \frac{u-d}{u}, \tag{11}$$

$$\nabla^2 F(u) = L_u + L_u' u + \frac{\lambda d}{u^2}. \tag{12}$$

Here and henceforth, arithmetic operations involving vectors is componentwise, and where appropriate we identify a vector with the diagonal matrix having the same entries. Then the iteration for Newton's method is:

$$u_{n+1} = u_n - t_n \nabla^2 F(u_n)^{-1} \nabla F(u_n), \tag{13}$$

where the step size t_n is either chosen to be 1 or by means of a line search. The rapid convergence that is possible with Newton's method relies upon the Hessian of F not varying too rapidly. However, that F is even C^2 relies on the addition of ϵ. A small value of ϵ, necessary for edge preservation, will both slow down the convergence of Newton's method, and make the Hessian ill-conditioned, making the computation of (13) more difficult.

As an alternative, we adopt a generalization of the approach in [2]. At each iteration, we approximate $\nabla F(u)$ with a linear function G_n by substituting some of the terms with their value from the preceding iteration:

$$G_n(u) = L_{u_n} u + \lambda \frac{u-d}{u_n}. \tag{14}$$

The derivative of this approximation is simply

$$H_n = L_{u_n} + \lambda / u_n. \tag{15}$$

We use this approximate Hessian in a quasi-Newton iteration:

$$u_{n+1} = u_n - t_n {H_n}^{-1} \nabla F(u_n). \tag{16}$$

In the case of $t_n \equiv 1$, this can be reformulated as

$$u_{n+1} = H_n^{-1} \frac{\lambda d}{u_n}, \tag{17}$$

which shows that (16) is equivalent to solving the linear equation $G_n(u_{n+1}) = 0$.

Each H_n is positive semidefinite and sparse. Provided the components of each u_n are neither too large nor too small, each H_n will be well-conditioned and not expensive to invert. The approximate Hessian differs from the true one even when $n \to \infty$. Therefore, we cannot expect quadratic convergence. However, numerical results show that the algorithm still converges to a correct solution much faster than gradient descent, as we will see in Section V.

A. A quasi-Newton method for general noise

We now describe the extension of (16) to an algorithm that is flexible in the choice of data fidelity term, for use in removing any of many different types of noise. We wish to solve the following:

$$\min_u F(u) = \int |\nabla u| + \lambda \int D(u,d). \tag{18}$$

In order to extend the proposed algorithm for a more general data fidelity term, we model the approximation of the Hessian

in a similar way as we did for the Poisson noise data fidelity term. To do that, we require that $\nabla_u D(u,d)$ can be written in the form of $E_1(u,d)u - E_2(u,d)d$. Then we approximate the Hessian with $H_n(u) = L_{u_n}(u) + E_1(u_n)I$, and obtain a quasi-Newton iteration as in (16).

As an example, we consider the L^1 data fidelity term, for which $D(u,d) = |u-d|$. As before, to obtain differentiability we will use

$$D(u,d) = \sqrt{(u-d)^2 + \delta}, \tag{19}$$

for some small $\delta > 0$. Note that the solution u_δ of (18) will converge in L^1 to u_0 as $\delta \to 0$. We have

$$\nabla_u D(u,d) = \frac{1}{\sqrt{(u-d)^2+\delta}}u - \frac{1}{\sqrt{(u-d)^2+\delta}}d, \tag{20}$$

so $E_1(u,d) = E_2(u,d) = ((u-d)^2+\delta)^{-1/2}$.

IV. Proof of convergence

Theorem 4.1: Let the functional F be defined on $E \subset \mathbb{R}^N$ by (18). Fix $d \in E$. Let $u_0 \in E$, and let S be the sublevel set $S = \{u \in E : F(u) \le F(u_0)+1\}$. Assume that $u \mapsto D(u,d)$ is both C^2 and strictly convex on S, and weakly coercive. Suppose that $\nabla_u D(u,d) = E_1(u,d)u - E_2(u,d)d$, and assume that $E_1 \ge mI$ uniformly on S, for some $m > 0$. Let u_n be defined iteratively by (16), with the step size t_n chosen by a backtracking line search. Then u_n converges to the unique minimizer of F.

The restriction to a subset E allows the possibility of $D(u,d)$ not being defined on all of $\mathbb{R}^N$. For example, in the case of the Poisson data fidelity term, we would take E to be the positive orthant of $\mathbb{R}^N$. We also point out that in practice, the line search is usually not required, and the algorithm converges with a uniform step size of 1. *Proof:*

We proceed inductively. Note that $u_0 \in \operatorname{int} S$. Given $u_n \in \operatorname{int} S$, as in Section III we let $H_n = L_{u_n} + E_1(u_n)$. Since L_{u_n} is positive semidefinite, $H_n \ge mI$ uniformly on S, and in particular is invertible. The boundedness on $\mathbb{R}^N$ of the finite difference operators D_x, D_y ensures that F is C^2. The weak coercivity of $u \mapsto D(u,d)$, and hence of F, guarantees that S is compact. Then H_n and $\nabla^2 F$ are bounded above on S, by some constants M_1 and M_2 respectively.

Our quasi-Newton step direction is $v_n = -H_n^{-1}\nabla F(u_n)$. We choose a step size t_n with a backtracking line search as follows. We fix $\alpha \in (0, \frac{1}{2})$, $\beta \in (0,1)$. Starting with $t_n = 1$, we replace t_n with βt_n until the following exit condition is satisfied:

$$\begin{aligned} F(u_n + t_n v_n) &\le F(u_n) + \alpha t_n \nabla F(u_n)^T v_n \\ &= F(u_n) - \alpha t_n \nabla F(u_n)^T H_n^{-1} \nabla F(u_n). \end{aligned} \tag{21}$$

To show that this condition is eventually met, suppose $t \le t^* := \min\{\frac{2m}{M_2}(1-\alpha), 1\}$ is chosen sufficiently small that $u_n + tv_n \in \operatorname{int} S$. (We shall show shortly that the former condition implies the latter.) Then by Taylor's theorem, for some ξ between u_n and $u_n + tv_n$ (hence belonging to S, by the convexity of F):

$$\begin{aligned} F(u_n + tv_n) - F(u_n) &= tv_n^T \nabla F(u_n) + t^2 v_n^T \frac{\nabla^2 F(\xi)}{2} v_n \\ &\le tv_n^T \nabla F(u_n) + t^2 \frac{M_2}{2} \|v_n\|^2 \\ &= -t\nabla F(u_n)^T H_n^{-1} \nabla F(u_n) \\ &\quad + t^2 \frac{M_2}{2} \left\| H_n^{-1} \nabla F(u_n) \right\|^2 \\ &\le -t\nabla F(u_n)^T H_n^{-1} \nabla F(u_n) \\ &\quad + t^2 \frac{M_2}{2m} \nabla F(u_n)^T H_n^{-1} \nabla F(u_n). \end{aligned} \tag{22}$$

Since $t(-1 + tM_2/2m) \le -\alpha t$ by the choice of t, the exit condition (21) is satisfied. Since H_n is positive definite, H_n^{-1} is too, so $F(u_n + tv_n) \le F(u_n)$ whenever $t \le t^*$.

Let $J = \{t \ge 0 : u_n + tv_n \in \operatorname{int} S\}$. Since S is convex, J is an interval. Let $t_S = \sup J$. If $t_S = \infty$, then $t^* \in J$. Otherwise, by the continuity of F, there must be $t_J \in J$ such that $F(u_n + t_J v_n) > F(u_n)$. Then $t_J > t^*$ by the preceding calculation. It follows in either case that $t \le t^* \Rightarrow t \in \operatorname{int} S$.

Let t_n denote the first t satisfying the exit condition (21) produced by the line search, so $t_n \ge \beta t^*$. Let $u_{n+1} = u_n + t_n v_n$. We then have that $F(u_n)$ is a decreasing sequence, and converges to some ℓ. Since (u_n) is bounded (as S is), it suffices to show that every convergent subsequence of (u_n) converges to a minimizer of F, since the strict convexity of F guarantees that there will be at most one minimizer.

Suppose $u_{n_k} \to u^*$. Then both $F(u_{n_k})$ and $F(u_{n_{k+1}})$ converge to ℓ. By the backtracking exit condition, we have

$$\begin{aligned} \|F(u_{n_k+1}) - F(u_n)\| &\ge \alpha t_{n_k} \nabla F(u_{n_k})^T H_{n_k}^{-1} \nabla F(u_{n_k}) \\ &\ge \alpha\beta t^* \frac{1}{M_1} \|\nabla F(u_{n_k})\|^2, \end{aligned} \tag{23}$$

from which follows that $\nabla F(u_{n_k}) \to 0$. ∇F is continuous, so $\nabla F(u^*) = 0$. This and the convexity of F imply that u^* is a minimizer of F, completing the proof. ∎

We now check that the Poisson and (approximate) L^1 data fidelity terms meet the conditions of the theorem. For $D(u,d) = u - d\log u$, we let $E = \{u \in \mathbb{R}^N : \text{ each } u_i > 0\}$, and fix any $d, u_0 \in E$. Then $u \mapsto D(u,d)$ is clearly C^2 and weakly coercive on all of E. Since $\nabla_u^2 D(u,d) = (d/u^2)$, we have strict convexity on E as well. We have $E_1(u,d) = 1/u$. The condition that $E_1 \ge mI$ for some $m > 0$ uniformly would not be true on E, but it is true on the sublevel set S. Indeed, $\|u\|$ is bounded on S, so $1/\|u\|$ is bounded away from 0.

In the case of $D(u,d) = \sqrt{(u-d)^2+\delta}$, we can take $E = \mathbb{R}^N$. In this case we have strong coercivity of $u \mapsto D(u,d)$. Since $\nabla_u^2 D(u,d) = \delta\big((u-d)^2+\delta\big)^{-1/2}$, both strict convexity and being C^2 are true for $\delta > 0$. Since $E_1 = \big((u-d)^2+\delta\big)^{-1/2}$, that $E_1 \ge mI$ uniformly on S follows from the coercivity, as in the Poisson case.

V. Numerical results

The numerical results achieved through this quasi-Newton method show that the algorithm works quite well. First, we

present reconstruction of images corrupted by Poisson noise. As a test image we use an image consisting of three concentric circles on a black background with a dark frame (Figure 1(a); the colormap has been adjusted to emphasize contrast between different intensity values). It contains both constant regions and sharp edges, a case in which total variation is an appropriate regularization term. Poisson noise is added to the image, which produces a signal dependent corruption. A one-dimensional slice through the center of the noisy image is shown in Figure 1(d). For an initial point u_0 in our iteration we

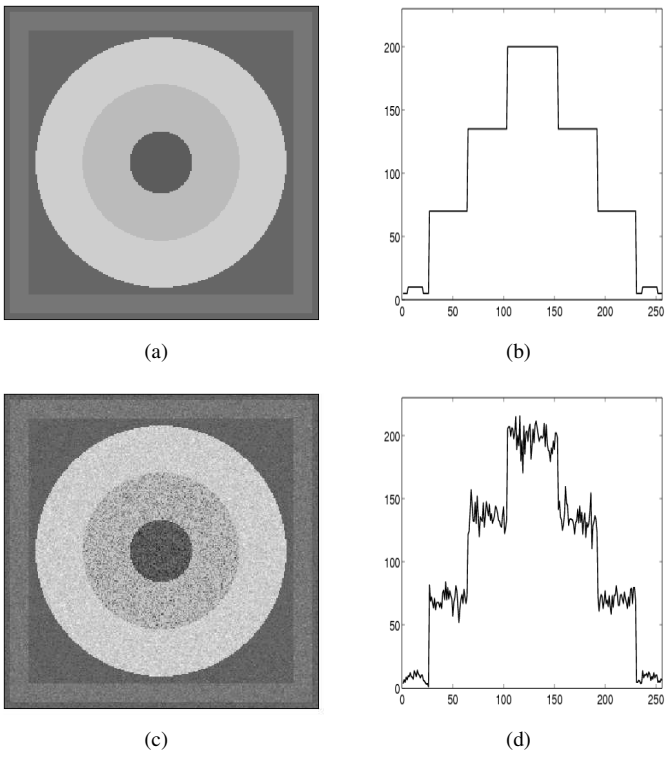


Fig. 1. (a) An image consisting of uniform regions with sharp edges. (b) A cross-section of the image through the center showing the intensity pattern we want to preserve. (c) The image corrupted by Poisson noise. (d) The cross-section indicating that the noise is signal dependent: higher intensities have greater variance.

use the noisy image. We choose λ according to the discrepancy principle, so that the value of the data fidelity term of the reconstructed image is the same as that of the original image.

First we apply the quasi-Newton method with the L^2-norm data fidelity term (which is the algorithm of [2], applied to the Rudin-Osher-Fatemi model [4]) to denoise the image. This method has a fast convergence rate, but it implicitly makes the assumption that the noise is Gaussian, which does not account for the signal dependence of the Poisson noise. Thus, equal denoising is applied to stronger and weaker noise, which results in underregularizing or oversmoothing. It can be seen in Figure 2(b) how the obtained solution still has significant noise in the region of higher intensity values. If we decrease the regularization parameter only enough to remove the noise, we are unable to correctly reconstruct the image in the regions of lower intensity values: the frame is merged with the background (Figure 2(d)).

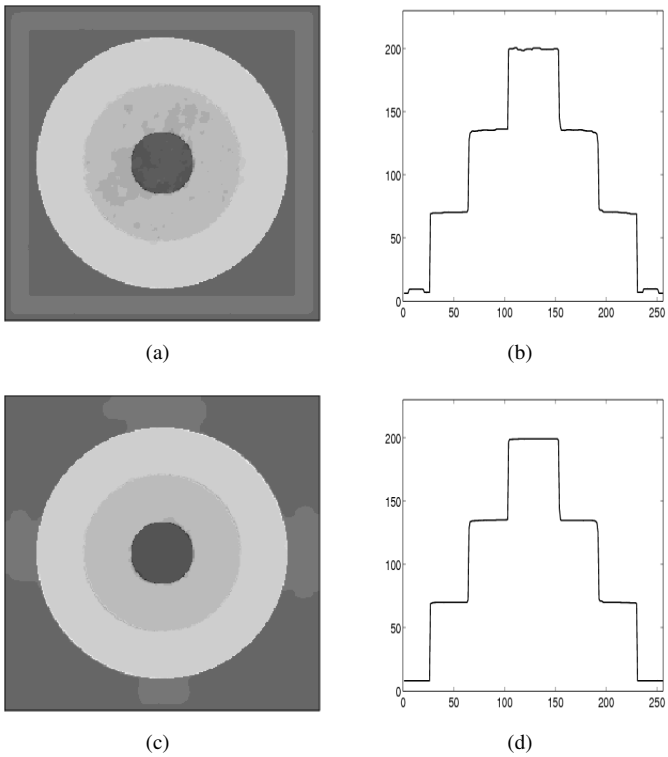


Fig. 2. (a),(b) A solution obtained through a quasi-Newton method with an L^2 data fidelity term: the edges are preserved correctly, however some noise is still present in the image. (c),(d) The regularization parameter is decreased: the cross-section is completely smoothed, cleaned from any noise, but the frame is missing.

Using a Poisson noise data fidelity term eliminates this problem. After only 20 iterations, the algorithm converges to the solution displayed in Figure 3. The noise is removed at all scales, and the low-contrast frame is well preserved.

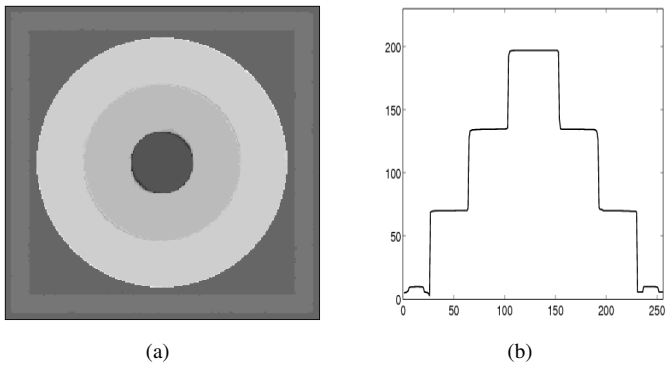


Fig. 3. (a) The result obtained with the Poisson noise data fidelity term. (b) The cross-section showing the preservation of the sharp edges and complete removal of noise. The frame is much clearer than in the noisy image.

For a second example, we use the standard cameraman image (Figure 4). Salt-and-pepper noise is added, with 20% of the pixels corrupted. As before, we compare with the Rudin-Osher-Fatemi model. Figure 4(c) shows the result, where the parameter λ is chosen to give the weakest regularization that removes the noise. The image is oversmoothed. If the L^1-norm data fidelity term is used instead, we obtain Figure 4(d).

The noise is removed, and the image is well restored. The algorithm converged in 40 iterations, where δ was initially chosen to be relatively large, $\delta = 1$, to speed convergence, then progressively decreased to $\delta = 10^{-6}$ to improve accuracy.

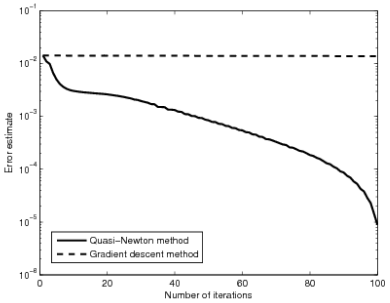
(a) (b)

(c) (d)

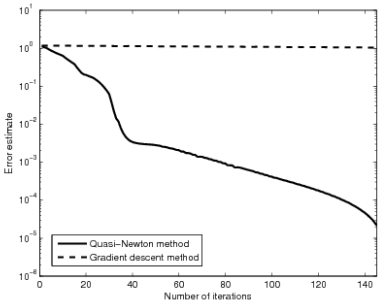
Fig. 4. (a) Cameraman image. (b) 20% of the pixels have been corrupted by salt-and-pepper noise. (c) Result of denoising with an L^2-norm data fidelity term. The weakest regularization that removes the noise still oversmooths the image badly. (d) Reconstruction with an L^1-norm data fidelity term. The noise is removed, and the image is well restored.

A. Convergence rate

Figure 5(a) compares the convergence rate of a gradient descent method and our quasi-Newton method. To estimate the error we consider the quantity $||u_n - u^*||/||u^*||$, where u^* is the minimum achieved through the quasi-Newton algorithm. The gradient descent method with a step size of 10^{-4} converges very slowly toward the solution. The first 100 iterations are displayed; the rate remains the same afterwards, and after 1000 iterations the error has changed slightly from 1.43×10^{-2} to 1.23×10^{-2}. The quasi-Newton method, on the other side, reaches an error of 10^{-5} within the first 100 iterations. It turns out the line search is not necessary for the quasi-Newton method with the noisy image as an initial guess: even if it is incorporated in the algorithm, the step size chosen is still 1. However, it is required when the starting point is far from the correct solution. We test the algorithm choosing a random image for an initial guess. In this case the gradient descent method requires more than 5000 iterations to converge. The quasi-Newton method uses the backtracking line search to choose a correct step size. First it selects a small step until it gets closer to the solution, and then quickly converges with a step size of 1. As a result the algorithm achieves global convergence with an error 10^{-5} after 40 more iterations (Figure 5(b)).



(a)



(b)

Fig. 5. (a) Convergence rate with a noisy image as a starting point. The quasi-Newton method starts with rapid convergence to a solution, visually indistinguishable from the correct one, and then gradually reduces the error to 10^{-5}. The gradient descent method approaches the minimum at a constant rate and requires much more than 100 iterations to find the solution. (b) We use a random image as a starting point. The randomness in the initial choice does not affect significantly the convergence rate of the quasi-Newton algorithm and it converges in slightly more iterations.

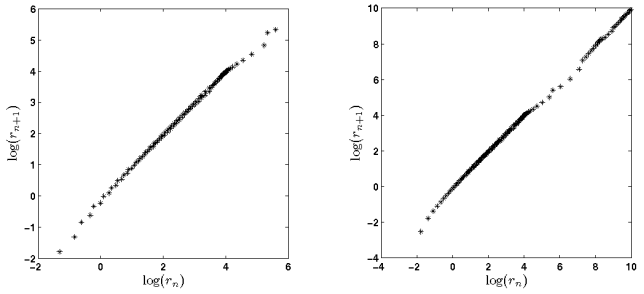
To analyze the rate of convergence, we plot the error of each iteration versus the error of the previous iteration on a log-log plot (Figure 6). The graph approximates a line with a slope 1, which suggests a linear convergence rate. Assuming that the relationship is indeed linear, we can calculate the convergence rate and for the first several iterations it is as low as 0.8, and is always lower than the rate of 0.9998 for the gradient descent method. The huge difference between the number of iterations required to find a solution justifies the higher computational cost for computing the approximate Hessians. The image examples clearly show how the proposed regularization removes the noise, and at the same time preserves sharp edges and low-contrast features. So we conclude that the proposed algorithm is very appropriate and efficient for regularization of images corrupted by non-Gaussian noise.

REFERENCES

[1] T. Le, R. Chartrand, and T. J. Asaki, "A variational approach to reconstructing images corrupted by Poisson noise," *J. Math. Imaging Vision*, 2007. To appear.

[2] C. R. Vogel and M. E. Oman, "Iterative methods for total variation denoising," *SIAM J. Sci. Comput.*, vol. 17, no. 1, pp. 227–238, 1996.

[3] D. Dobson and C. Vogel, "Convergence of an iterative method for total variation denoising," *SIAM J. Numer. Anal.*, vol. 34, pp. 1779–1791, 1997.



(a) (b)

Fig. 6. Log-log graphs of the relative residual r_{n+1} versus r_n, where $r_n = \|u_n - u^*\| / \|u^*\|$, are used to demonstrate the convergence rate of the quasi-Newton algorithm: (a) with the noisy image as a starting point; and (b) with a random image as a starting point. In both cases the graph implies a linear convergence rate.

[4] L. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms," *Physica D*, vol. 60, pp. 259–268, 1992.

[5] M. Green, "Statistics of images, the TV algorithm of Rudin-Osher-Fatemi for image denoising and an improved denoising algorithm," CAM Report 02-55, UCLA, 2002.

[6] E. Jonsson, S.-C. Huang, and T. Chan, "Total variation regularization in positron emission tomography," Tech. Rep. 98-48, UCLA Group in Computational and Applied Mathematics, 1998.

[7] M. Nikolova, "Minimizers of cost-functions involving nonsmooth data-fidelity terms. Application to the processing of outliers," *SIAM J. Numer. Anal.*, vol. 40, pp. 965–994, 2002.

[8] M. Nikolova, "A variational approach to remove outliers and impulse noise," *J. Math. Imaging Vision*, vol. 20, pp. 99–120, 2004.